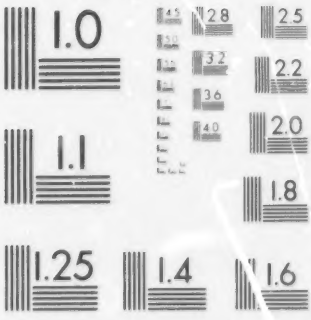

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

DOCUMENT RESUME

ED 257 838 TM 850 136

AUTHOR	Subkoviak, Michael J.
TITLE	Tables of Reliability Coefficients for Mastery Tests.
PUB DATE	Mar 85
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Speeches/Conference Papers (150) -- Statistical Data (110)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Estimation (Mathematics); *Mastery Tests; Statistical Studies; *Tables (Data); *Test Interpretation; *Test Reliability
IDENTIFIERS	Kappa Coefficient

ABSTRACT

Current methods of obtaining reliability coefficients for mastery tests are laborious from a practitioner's perspective. Some methods require two test administrations; while others require access to computer facilities and/or advanced measurement and statistical procedures. This report provides tables from which practitioners can read such reliability coefficients directly. The method used to construct the tables is reviewed. Comments on the accuracy of the tabled values are included. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED257838

Tables of Reliability Coefficients for Mastery Tests

Michael J. Subkoviak

University of Wisconsin-Madison

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1985.

TM 850 136

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. Subkoviak

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

From a practitioner's perspective, current methods of obtaining reliability coefficients for mastery tests are quite laborious. For example, some methods demand two test administrations; while others require access to computer facilities and/or involve advanced measurement and statistical procedures. Thus, the present paper provides tables from which practitioners can read such reliability coefficients directly. The method used to construct the tables is reviewed; and comments on the accuracy of the tabled values are included.

Tables of Reliability Coefficients for Mastery Tests

Methods for obtaining reliability estimates for mastery tests can be quite laborious from a practitioner's point of view. For example, the method proposed by Swaminathan, Hambleton, and Algina (1974) requires two administrations of the same or parallel tests. Given examinees' scores on both administrations and the cutoff score which distinguishes masters from nonmasters, two different reliability indices can be computed: (1) the agreement coefficient and (2) the kappa coefficient.

The agreement coefficient is simply the proportion of examinees consistently classified as masters or as nonmasters on both administrations. When the mastery-nonmasterly classifications on the two administrations are summarized as in Table 1, the agreement coefficient, designated p_o, is given by

$$p_o = p_{11} + p_{22} , \qquad (1)$$

where p_{11} and p_{22} are the proportions of examinees classified, respectively, as masters and nonmasters on both administrations. The upper bound of the agreement coefficient is 1.00, which occurs if classifications on both administrations are consistent for all examinees in the group. When the

---

Insert Table 1 about here

---

two administrations in Table 1 involve the same or parallel tests, the lower bound of the agreement coefficient is given by

$$p_{chance} = (p_{11} + p_{12})(p_{11} + p_{21}) + (p_{21} + p_{22})(p_{12} + p_{22}) , \qquad (2)$$

where p_{chance} represents the expected proportion of consistent classifications when there is no relationship between outcomes on the two test administrations (Huynh, 1978).

The aforementioned kappa coefficient, designated κ, is given by

$$\kappa = (p_o - p_{chance})/(1 - p_{chance}) , \qquad (3)$$

where p_o and p_{chance} are obtained from (1) and (2). As such, kappa reflects the proportion of consistent classifications beyond that expected by chance. The upper and lower bounds of kappa are 1.00 and 0, which occur, respectively, when there is perfect agreement and no relationship between outcomes on the two test administrations.

Computer methods for estimating the agreement and kappa coefficients from a single test administration have been proposed, thereby eliminating the need for a second test administration (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976). However, these methods are also difficult for practitioners to implement; since they require access to computer facilities and appropriate software, and they assume a somewhat advanced background in test theory.

Approximation methods involving hand calculations of the agreement and kappa coefficients from a single test administration have also been proposed (Huynh, 1976, p. 258; Peng & Subkoviak, 1980, p. 363). While these methods are the simplest thus far proposesd, they still involve the use of statistical tables of the bivariate and univariate normal distributions, which may not be entirely familiar to or readily available to practitioners. Thus, the present paper provides even greater simplicity: tables from which practitioners can

directly read approximate values of the agreement coefficient or the kappa coefficient.

Tables of Agreement and Kappa Coefficients

Table 2 contains approximate values of the agreement coefficient, and Table 3 contains approximate values of the kappa coefficient.

Insert Tables 2 and 3 about here

In order to use either table, two values are needed, which can be obtained from the data for a single test administration: (1) the norm-referenced reliability of the test (r) and (2) the cutoff score of the test expressed as a standard score (z).

The norm-referenced reliability coefficient r can be computed using well-known and widely published formulae (Stanley, 1971); some of the more common indices of this type are the Kuder-Richardson 20 and 21 coefficients, Cronbach's alpha coefficient, and Hoyt's reliability coefficient. For example, Kuder-Richardson formula 21 provides practitioners with a very simple means of estimating r:

$$r_{KR-21} = \frac{nS^2 - M(n - M)}{(n - 1)S^2}, \qquad (4)$$

where n is the number of test items, M is the mean of the scores, and S^2 is the variance of the scores. Formula 4 is appropriate for test items scored as right or wrong; and it generally provides underestimates of reliability coefficient r, which lead to conservative estimates of the agreement and kappa coefficients in Tables 2 and 3. If items are not binary

scored or if less conservative estimates are desired, one of the other formulae noted previously for estimating r can be employed.

The standard score z, which appears in Tables 2 and 3, is obtained as follows:

$$z = \frac{(c - .5 - M)}{S}, \qquad (5)$$

where c is the raw cutoff score of the test, M is the mean of the scores, and S is the standard deviation of the scores. The value .5 in Equation 5 is a correction for continuity which arises from the fact that Tables 2 and 3 were obtained by approximating the discrete test score distribution with the continuous normal distribution, as discussed later. The computed value of z given by Equation 5 may be either negative or positive. However, due to the symmetry of the approximating normal distribution, a negative z value like $-.10$ will lead to the same agreement or kappa coefficient as a positive z value like $+.10$. Thus, the unsigned or absolute value $|z|$ is sufficient in order to make use of Tables 2 and 3.

An Example

The use of Tables 2 and 3 will be illustrated employing a set of real data, which is described in greater detail elsewhere (Subkoviak, 1980). A 10 item multiple-choice test, with a cutoff score of 8, was administered to $N = 30$ students. The mean of the test was $M = \Sigma x/N = 4.63$, and the variance was $S^2 = [\Sigma x^2/(N - 1)] - [(\Sigma x)^2/N(N - 1)] = 3.27$. Using Equation 4, the reliability of the test was $r_{KR-21} = [nS^2 - M(n - M)]/[(n - 1)S^2] = [(10)(3.27) - (4.63)(10 - 4.63)]/[(10 - 1)(3.27)] = .27$, or $r_{KR-21} = .30$

approximately. Using Equation 5, the z value was $z = (c - .5 - M)/S = (8 - .5 - 4.63)/\sqrt{3.27} = 1.59$, or $z = 1.60$ approximately. Entering Table 2 with $r = .30$ and $|z| = 1.60$, it can be seen that the coefficient of agreement is $p_o = .91$, approximately. Similarly, the kappa coefficient provided in Table 3 is $\kappa = .10$, approximately. The values of the agreement and kappa coefficient are quite different ($p_o = .91$ vs. $\kappa = .10$) because the two coefficients are distinct measures of consistency--a point discussed in greater detail in the next section of the paper. Since $r = .27$ and $|z| = 1.59$ in the example, somewhat more precise estimates of p_o and κ could be obtained from Tables 2 and 3 by interpolation (Subkoviak, 1980, pp. 141-142); but for practical purposes, the slight gain in precision may not be worth the additional effort.

Tables 2 and 3 can also be used to determine the agreement and kappa coefficient of a test that has been lengthened or shortened by a factor of ℓ. Suppose one wished to determine in the previous example what the agreement and kappa coefficient would be if 5 more items, equivalent to the original 10, were added to the test. Since the lengthened test of 15 items is 1.5 times the original length, ℓ would equal 1.5. The mean of the lengthened test would be $M_\ell = \ell M = (1.5)(4.63) = 6.95$; the cutoff of the lengthened test would be $c_\ell = \ell c = (1.5)(8) = 12$; and the variance of the lengthened test would be $S^2_\ell = \ell S^2[1 + (\ell - 1)r] = (1.5)(3.27)[1 + (1.5 - 1)(.27)] = 5.57$ (Lord & Novick, 1968, p. 86). Substituting these values into Equation 5 produces the result $z_\ell = (c_\ell - .5 - M_\ell)/S_\ell = (12 - .5 - 6.95)/\sqrt{5.57} = 1.93$, or $z = 1.90$ approximately. Finally, the reliability of the lengthened test would be $r_\ell = \ell r/[1 + (\ell - 1)r] = (1.5)(.27)/[1 + (1.5 - 1)(.27)] = .36$, or $r_\ell = .40$ approximately. Entering Tables 2 and 3 with $r_\ell = .40$ and $|z_\ell| = 1.90$, the agreement and kappa coefficients of the lengthened test

are $p_o = .95$ and $\kappa = .12$, approximately, which are slightly larger than the original values ($p_o = .91$ and $\kappa = .10$); since lengthening a test increases its reliability.

Discussion

It may be noted that corresponding entries in Tables 2 and 3 are generally quite different, as in the example where $p_o = .91$ and $\kappa = .10$. Such differences are due to the fact that the agreement coefficient and the kappa coefficient are distinct measures of consistency (see Subkoviak, 1980, pp. 152-154). The agreement coefficient is the total proportion of consistent classifications on two test administrations; whereas the kappa coefficient reflects the proportion of consistent classifications, beyond that expected by chance. In concrete terms, what this means is that the kappa coefficient is more sensitive than the agreement coefficient to changes in test reliability r, as can be seen by comparing corresponding rows of Tables 2 and 3; and kappa is less sensitive to changes in $|z|$, which reflect the location of the cutoff within the distribution of scores, as can be seen by comparing corresponding columns of Tables 2 and 3. An awareness of these differences is important when interpreting and reporting values of the two coefficients.

The question of what is an acceptable value of an agreement or kappa coefficient naturally arises when interpreting and reporting obtained values of these indices. Consider the coefficient of agreement (p_o), which can be thought of as the probability that a randomly selected examinee will be consistently classified on two test replications. The question of how large this probability value should be depends upon the seriousness of the decisions being made with the test results. If the test is being used to decide who

will graduate and who will not, this probability should be quite large, perhaps .95, as might occur in Table 2 with a published test having reliability $r = .90$ and standardized cutoff $z = -1.50$ (a standard that implies about 7% of a normally distributed group score below the cutoff). On the other hand, if the test results are being used to make routine classroom decisions like who will move-on to the next unit of instruction and who will remain on the present unit, the probability can be somewhat lower, perhaps .85, as might occur in Table 2 with a teacher-made test having reliability $r = .70$ and standardized cutoff $z = -1.00$ (implying about 16% of a normally distributed class score below the cutoff).

The question of what constitutes an acceptable value of a kappa coefficient can best be answered by first reviewing what the coefficient measures. The formula for kappa (3) involves the values p_{chance}, p_o, and 1.0 which represent, respectively, the probability of consistent classification when no relationship exists, the observed relationship exists, and a perfect relationship exists between the outcomes on two test administrations. Therefore, the numerator of kappa $(p_o - p_{chance})$ reflects the gain in consistency between the no relationship condition (p_{chance}) and the observed relationship (p_o); and the denominator $(1 - p_{chance})$ reflects the maximum gain in consistency possible between the no relationship condition (p_{chance}) and the perfect relationship condition (1.0). Thus, $\kappa = (p_o - p_{chance})/(1 - p_{chance})$ is a ratio of the actual gain in consistency due to the test to the maximum gain possible; or in other words, kappa reports the actual contribution the test makes to consistency as a proportion of the maximum possible contribution that could be made. Kappa, therefore, is a measure of the extent to which a test is performing up to the maximum possible limit; and one normally expects more, in this sense, of

published tests than teacher-made tests. For example, published tests should probably be expected to have kappa values of .50 or greater, as would occur in Table 3 for a published test having reliability $r = .90$ and standardize cutoff $|z|$ between .00 and 2.00. On the other hand, teacher-made tests might be expected to have kappa values of .25 or greater, as would occur in Table 3 for a classroom test having reliability $r = .70$ and standardized cutoff $|z|$ between .00 and 2.00. However, notice that a test may not be living up to expectation in terms of the kappa coefficient; and yet the overall probability of consistent classification in terms of the agreement coefficient may still be acceptable. For example, if a published test has reliability $r = .80$ and standardized cutoff $|z| = 2.00$, the associated kappa value in Table 3 is $\kappa = .42$, which is below the .50 benchmark previously suggested for a standardized test; yet the associated agreement coefficient in Table 2 is $p_o = .97$, which is above the .95 benchmark previously suggested for tests used to make important decisions. This is one more illustration of the fact that the agreement and kappa coefficients are distinct measures of consistency, requiring individual interpretation.

Construction of the Tables

Tables 2 and 3 were constructed using a procedure proposed by Peng and Subkoviak (1980, p. 363) for estimating the agreement or kappa coefficient. This procedure assumes that <u>if</u> two test administrations were actually conducted, the joint distribution of scores on the two testings could be approximated by a bivariate normal distribution. Under this assumption, the agreement and kappa coefficient are, respectively, given by

$$p_o = 1 + 2(p_{zz} - p_z) \qquad (6)$$

$$k = \frac{p_{zz} - p_z^2}{p_z - p_z^2}; \qquad (7)$$

where z is the cutoff expressed as a standard score (see Equation 5): p_z is the probability that a standard normal variable is less than value z: and p_{zz} is the probability that two standard normal variables, having correlation r (see Equation 6), are less than z.

Tables 2 and 3 were obtained from Equations 6 and 7 by first specifying values for z and r and by then determining the corresponding values of p_z and p_{zz} in (6) and (7), which can be obtained by computer routines or from tables of the univariate and bivariate normal distribution. For example, the first entry in Tables 2 and 3 was obtained by specifying the values $z = .00$, $r = .10$ and determing the corresponding probabilities $p_z = .5000$, $p_{zz} = .2659$ from the standard univariate and bivariate normal distributions. Substituting these values into Equation 6 provides the agreement coefficient $p_o = 1 + 2(p_{zz} - p_z) = 1 + 2(.2659 - .5000) = .5318$ or $p_o = .53$, approximately; and Equation 7 provides the kappa coefficient $\kappa = (p_{zz} - p_z^2)/(p_z - p_z^2) = (.2659 - .5000^2)/(.5000 - .5000^2) = .0636$ or $.06$, approximately. All other entries in Tables 2 and 3 were obtained in the same way.

Peng and Subkoviak (1980) found that Equations 6 and 7, on which Tables 2 and 3 are based, generally provide good approximations even when the test data are <u>not</u> normally distributed. They simulated nonnormal data for 125 different conditions, including U-shaped, uniform, platykurtic, leptokurtic, and skewed test score distributions; and they then compared the <u>exact</u>

agreement and kappa coefficients for the data to approximations of these two coefficients given by Equations 6 and 7. The average discrepancy between exact and approximate values over all 125 conditions was .013 for the agreement coefficient and .037 for the kappa coefficient. As would be expected, the greatest discrepancies occurred for the most nonnormal distributions of test scores, which were U-shaped; the average discrepancy over 25 such cases was .019 for the agreement coefficient and .043 for the kappa coefficient. As the simulated test score distributions become more near-normal, discrepancies between exact and approximate values decreased; for example, the average discrepancy over 25 leptokurtic cases was .008 for the agreement coefficient and .036 for the kappa coefficient. Thus, it appears that Tables 2 and 3 should generally provide practitioners with useful approximations of the agreement and kappa coefficients over a variety of realistic data conditions and with minimal effort.

For purposes of completeness it might be noted that tables of agreement and kappa coefficients have also been produced by Huynh for short tests containing between five and ten items; and these tables are reproduced in Subkoviak (1980). However, the Huynh tables are based on the assumption that the test data follow a beta-binomial distribution rather than a normal distribution, as assumed in Tables 2 and 3.

References

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Huynh, H. (1978). Reliability of multiple classifications. Psychometrika, 43, 317-325.

Lord, [illegible], & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley Publishing Company.

Marshall, J. L., & Haertel, E. H. (1976). The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, University of Wisconsin-Madison.

Peng, C.-Y. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. Journal of Educational Measurement, 17, 359-368.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education.

Subkoviak, M. J. (1976). Estimating reliability from a single adminstration of a mastery test. Journal of Educational Measurement, 13, 265-276.

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. A. Berk (Ed.), Criterion-referenced measurement. Baltimore, MD: Johns Hopkins University Press.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 11, 263-267.

Table 1

Classification of Examinees on Two Test Administrations

		Admin 2		
		Master	Nonmaster	
Admin 1	Master	p_{11}	p_{12}	$(p_{11} + p_{12})$
	Nonmaster	p_{21}	p_{22}	$(p_{21} + p_{22})$
		$(p_{11} + p_{21})$	$(p_{12} + p_{22})$	

Table 2

Approximate Values of the Agreement Coefficient p_o

	r								
|z|	.10	.20	.30	.40	.50	.60	.70	.80	.90
.00	.53	.56	.60	.63	.67	.70	.75	.80	.86
.10	.53	.57	.60	.63	.67	.71	.75	.80	.86
.20	.54	.57	.61	.64	.67	.71	.75	.80	.86
.30	.56	.59	.62	.65	.68	.72	.76	.80	.86
.40	.58	.60	.63	.66	.69	.73	.77	.81	.87
.50	.60	.62	.65	.68	.71	.74	.78	.82	.87
.60	.62	.65	.67	.70	.73	.76	.79	.83	.88
.70	.65	.67	.70	.72	.75	.77	.80	.84	.89
.80	.68	.70	.72	.74	.77	.79	.82	.85	.90
.90	.71	.73	.75	.77	.79	.81	.84	.87	.90
1.00	.75	.76	.77	.77	.81	.83	.85	.88	.91
1.10	.78	.79	.80	.81	.83	.85	.87	.89	.92
1.20	.80	.81	.82	.84	.85	.86	.88	.90	.93
1.30	.83	.84	.85	.86	.87	.88	.90	.91	.94
1.40	.86	.86	.87	.88	.89	.90	.91	.93	.95
1.50	.88	.88	.89	.90	.90	.91	.92	.94	.95
1.60	.90	.90	.91	.91	.92	.93	.93	.95	.96
1.70	.92	.92	.92	.93	.93	.94	.95	.95	.97
1.80	.93	.93	.94	.94	.94	.95	.95	.96	.97
1.90	.95	.95	.95	.95	.95	.96	.96	.97	.98
2.00	.96	.96	.96	.96	.96	.97	.97	.97	.98

Table 3

Approximate Values of the Kappa Coefficient κ

	r								
|z|	.10	.20	.30	.40	.50	.60	.70	.80	.90
.00	.06	.13	.19	.26	.33	.41	.49	.59	.71
.10	.06	.13	.19	.26	.33	.41	.49	.59	.71
.20	.06	.13	.19	.26	.33	.41	.49	.59	.71
.30	.06	.12	.19	.26	.33	.40	.49	.59	.71
.40	.06	.12	.19	.25	.32	.40	.48	.58	.71
.50	.06	.12	.18	.25	.32	.40	.48	.58	.70
.60	.06	.12	.18	.24	.31	.39	.47	.57	.70
.70	.05	.11	.17	.24	.31	.38	.47	.57	.70
.80	.05	.11	.17	.23	.30	.37	.46	.56	.69
.90	.05	.10	.16	.22	.29	.36	.45	.55	.68
1.00	.05	.10	.15	.21	.28	.35	.44	.54	.68
1.10	.04	.09	.14	.20	.27	.34	.43	.53	.67
1.20	.04	.08	.14	.19	.26	.33	.42	.52	.66
1.30	.04	.08	.13	.18	.25	.32	.41	.51	.65
1.40	.03	.07	.12	.17	.23	.31	.39	.50	.64
1.50	.03	.07	.11	.16	.22	.29	.38	.49	.63
1.60	.03	.06	.10	.15	.21	.28	.37	.47	.62
1.70	.02	.05	.09	.14	.20	.27	.35	.46	.61
1.80	.02	.05	.08	.13	.18	.25	.34	.45	.60
1.90	.02	.04	.08	.12	.17	.24	.32	.43	.59
2.00	.02	.04	.07	.11	.16	.22	.31	.42	.58